A NEEDLE IN A HAYSTACK? HARNESSING ONOMATOPOEIA AND USER-SPECIFIC STYLOMETRICS FOR AUTHORSHIP ATTRIBUTION OF MICRO-MESSAGES

Antônio Theóphilo[⋆†]*, Luís A. M. Pereira*[⋆]*, Anderson Rocha*[⋆]

{antonio.theophilo, luis.pereira, anderson.rocha}@ic.unicamp.br
[⋆] Institute of Computing – University of Campinas
[†] CTI Renato Archer
Campinas/SP, Brazil

ABSTRACT

The world is facing a new era in which social media communication plays a fundamental role in people's lives. Along with irrefutable benefits, several collateral drawbacks have risen, one being the wide spread of false information with malicious intents, what is now commonly called "Fake News". The fight against this problem is not easy, especially when taking into account the nature of text messages involved on social media platforms (a sea of small messages and myriad users). In this work, we cope with the challenging problem of authorship attribution of small text messages posted on social media platforms. Differently from what has been done with longer texts, we rely upon data-driven approaches, exploiting recent advances of deep neural networks in the field of pattern recognition. By viewing small texts usually employed in social media as unidimensional signals, we devise modern deep-learning techniques tailored for this kind of data to find the author of these posts with promising results.

Index Terms— Authorship Attribution, Information Forensics, Deep Learning, Natural Language Processing, Social Media Data.

1. INTRODUCTION

The last years witnessed a significant growth of social media utilization by the world population. According to 2018 statistics, Twitter® accounts for more than 326 million active users [1] while for Facebook® this number exceeds 2.2 billion [2]. Social media platforms have radically reshaped the way people communicate and interact with each other, being used both to get in contact with friends and relatives as well as to share opinions with millions of people. We have gone from a model in which only large media corporations and governments were capable of reaching an audience of millions of individuals to another in which any citizen can produce content and spread messages to billions of people who can, in turn, interact with the author in unprecedented novel ways. The repercussions of this new scenario are diverse, being one of the major importance the fast people's ability to mobilize themselves over hot topics such as politics, injustices, and wars.

However, as almost every humankind technological achievement, social media platforms also gave rise to several undesirable and unplanned effects. To name a few, they have created or strengthened crimes like racism, misogyny, bullying (these often done anonymously), phishing, misinformation and even open doors to public opinion manipulation at a much grander scale. In a recent article published by The New York Times [3], an official Russian agency was accused of supporting a team of workers responsible for creating fictitious publications on social networks under false identities, in order to foster propaganda and disinformation campaigns aligned with this country's interests. According to Facebook® itself, more than 126 million people accessed the content generated by this organization, which also posted more than a thousand videos on YouTube® and more than 131,000 messages on Twitter® [4]. More recently, the same newspaper showed how these fake identities are created and nurtured so that they earn trustworthiness over the public [5]. In another article published by CNN, Davey-Attlee and Soares [6] presented in detail how social media platforms were manipulated in order to bias the 2016 American presidential election through the creation of invented profiles and false news, with the Brazilian 2018 general election being the more recent target.

There is a huge market involving the selling of false information through fake identities. Confessore et al. [7] show how rogue profiles are created and marketed, often using real data of innocent people, even names and images of minors. According to Varol et al. [8], nearly 15% of Twitter accounts (50 million) are automated profiles devised to simulate real people while for Facebook, as claimed by the company itself [9], this number exceeds 60 million accounts.

It is irrefutable the dangerous consequences that these platforms present to our modern society and the importance of verifying the authenticity of the information conveyed by them, in a way to avoid damages (sometimes irreversible) to the society at large. We are living in what some are calling the Fake News/Post-Truth Era, where mass communication platforms (as social media networks) are extensively being used to influence and deceive people. According to Gartner, by 2022, there will be more people in mature economies consuming false information than otherwise [10]. The end result is that trust on social networks by the general population has dramatically decreased over time. In a very fresh article published by Wired, Lapowsky [11] reveals a survey showing that only 37% of Californians trust on social media companies, despite tech industry being considered the most trustful industry in the same poll.

Social media networks can be seen as the primary source of individual text production nowadays. Despite their benefits regarding communication and information of people around the globe, much has been discussed on one of its lousy faces: the so-called "Fake News." A critical task in the fight against this disorder is to identify the author of the text message that conveys the misinformation. The

The research for this paper was financially supported by the São Paulo Research Foundation (FAPESP), DéjàVu grant #2017/12646-3 and grant #2018/10204-6; by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), DeepEyes grant; by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant #304472/2015-8 and by Google Latin America Research Awards (LARA 2018). The GPU cards used in this research were donated by NVIDIA Corporation through its GPU Grant Program.

mission of attributing the authority of a text to a person is called authorship attribution. This task consists of identifying the author of a text message taking into account exclusively the digital textual data produced, without considering another source of information, for example, the authorship claimed by the transporting media or any other meta-data.

One of the main challenges with authorship attribution for social media messages is obviously their length – too small when compared with other types of texts such as articles or romances. However, at the same time, there is information specific to these platforms that could be leveraged to identify writing patterns such as emojis, emoticons, hashtags, repetitions, and onomatopoeias.

Authorship attribution is definitely capable of helping in the identification of "Fake News" spread over social media platforms. It can also aid in the identification of false identities, especially when a single person pretends to be various fictitious ones [3] [6].

Here, we address the problem of micro-message authorship attribution as a unidimensional signal processing task. Unlike what has been done with longer texts, we propose a data-driven approach: we employ a convolutional neural network to extract features and perform the target classification directly from unidimensional signals built from small text messages.

In comparison to the current baseline, the contribution of our approach is the exploitation of the relative localization of the letters contained in the message, operating over an input space (manifold) with a dimension 130 times smaller and with almost no feature engineering. We experimented it over messages posted on Twitter® social network, and the results suggest our approach to be promising for authorship attribution, considering this new signal processing vantage point that treats micro-messages as unidimensional signals.

The rest of the text is organized as follows: Section 2 briefly discusses the related works in the area of authorship attribution and deep learning applied to textual tasks. Section 3 explains in details our proposed method to solve the problem while Section 4 unveils the performed experiments. In the end, Section 5 presents some conclusions and final remarks.

2. RELATED WORK

Previous work over authorship attribution have targeted large texts [12] [13], with some applying traditional machine learning classifiers over handcrafted features in a bag-of-words (BoW) approach [14] [15] [16] [17], analyzing the feature frequencies through the use of histograms and traditional shallow pattern classifiers (SVM - Support Vector Machines and RF - Random Forests, for instance). This approach has produced reasonable results with long texts because statistical data can be obtained from such big samples, even discarding some important ordering and localization information. Additional characteristics of this scenario that ease the classification task are the relatively narrow range of symbols used and the more formal parlance usually employed.

When considering the small texts published in social media platforms like Twitter ®, however, this approach is far from presenting reasonable results for real-world deployments [18]. The main reason is probably an untamed context that is extremely different from the one presented above, having the data the following attributes:

1. Individual samples with very small size, despite a considerable amount of data per user (thousands of tweets) and a huge amount of data in the platform as a whole;
2. A broader range of symbols used like non-ASCII Unicode characters, emojis, and emoticons;
3. A very wider parlance span, with the publishing of intentionally misspelled words, onomatopoeias and large use of Internet jargon.

Only recently, the problem of authorship attribution over very small messages, such as the ones posted on social media platforms, have been addressed. Rocha et al. [18] present an extensive bibliographic review and develop techniques that expand the state of the art for small texts without yet achieving results that could be used in real-world deployments. The accuracy for 50 authors is below 64% and for 1000 authors (a typical scenario in social media forensics) is under 44%, indicating an urgent demand for more promising solutions. Their approach, based on BoW and traditional machine-learning classifiers, besides losing locality information, leads to a high-dimensional model, resulting in a large memory footprint when dealing with scenarios of many messages and many possible authors, limiting its application.

In the past few years, deep learning techniques have started to present outstanding results in tasks over images [19], videos, and audio [20], treating them almost always as raw input signals. More recently, these techniques were also applied over textual media with amazing results [21] [22] [23] [24] [25] [26] [27] [28] [29], yet still not addressing the problem of authorship attribution, usually focusing over general semantic tasks like sentiment analysis, topic classification and language modelling, to name a few.

From all these previous works, Kalchbrenner et al. [25] presented a very interesting solution where convolutional neural networks techniques where applied for text sentence modeling and successfully evaluated it against three different tasks (movie review sentiment prediction, question classification, and Twitter® sentiment prediction). Their model, through an extension and generalization of the pooling operation (*Dynamic k-Max Pooling*), identifies the occurrence of the relevant n-grams for a task and also the relative position they appear. This model can be applied to hard-to-parse sentences, like tweets, since it does not rely on an externally provided parse tree. It generates a graph-based structure that can capture short and long-range relations between words that do not necessarily correspond to the syntactic relations of a traditional parse tree. Due to all these advantages, we have adapted their model to cope with the problem of authorship attribution of micro-messages as detailed in the next section.

3. METHODOLOGY

Authorship attribution has accomplished excellent results with the use of BoW approaches when applied over longer texts. Despite these achievements, in the case of small text messages usually posted in microblogs, we need to leverage every single bit of information and avoid approaches like histograms of features that, regardless of providing important frequency statistics from the associated data, also lose critical localization and ordering features.

Our approach to the problem consists of treating the small input text as a unidimensional input signal and feeding it to a convolutional neural network model to be classified into belonging to one of the suspect authors in a closed-set classification scenario. The rationale behind this strategy is to take the most of the ability of this kind of models to capture patterns specific to an author, leveraging every piece of data available in these small samples, including ordering information.

3.1. Input Signal Transformation

Each sample (tweet) is modeled into a sequence of character 4-grams due to the power of this representation in capturing important discriminating features [18]. We did not perform any normalization process (e.g., case converting or stemming) in the original data because we believe they would remove crucial information for our classification task. After learning all character 4-grams present in the training set, we noted that almost half of them occurs only once, bringing more noise than discriminative power to the model. At the same time, some 4-grams are used by all authors, with their occurrence not offering much help in classifying the authors beyond their ordering. We carried out experiments and realized that removing 4-grams that occur in all authors as well as the ones that take place only once does not compromise the accuracy and dramatically reduces the number of hyper-parameters in the embedding layer.

3.2. Network Architecture

Our approach is based on the work of Kalchbrenner et al. [25] in which the authors, besides proposing a convolutional neural network for different textual tasks, introduce a very interesting new operation called *dynamic k-max pooling* that allows the convolutional model to deal with samples of variable length and to capture short and long-range relations among elements.

The network is depicted in Figure 1 where a sentence is fed to the network and go through three steps: 1) a projection over the input space through a task-specific learned embedding layer; 2) a sequence of sets of feature maps computed in parallel (more details below) and; 3) a final fully connected layer followed by a softmax classifier.

The feature maps are, in turn, composed of a sequence of four basic layers/operations: 1) unidimensional convolution; 2) folding; 3) dynamic k-max pooling and; 4) non-linear activation (in our case the hyperbolic tangent function).

After the embedding layer, each sentence element is projected onto a multidimensional space. The convolution filters act over each dimension separately, being the folding operation the one which associates activations of adjacent dimensions in lower levels of the model, summing them up component-wise. Without this operation, these associations would only occur at the highest level of the model in the dense layer, probably losing or weakening relevant discriminative information. An alternative approach to folding could be an additional dimension over the convolutional filters at the expense of a significant increase in the number of parameters.

After the folding operation, the dynamic k-max pooling selects the k top activations from each dimension aiming to forward to the rest of the network more relevant data relative to the task in hand than would a regular max-pooling. This operation keeps the relative ordering of these selected activations, allowing the network to identify short and long-range relations among sentence elements. The dynamic term in the name refers to the fact that the value of k is fixed for the topmost pooling layer, but, for the other pooling layers, it varies with the layer position and the size of the sequence. For the lower pooling layer l, the value of k is given by the formula below where L is the total number of convolutional layers, s is the length of the sentence and k_{top} is the fixed value of k for the topmost pooling layer.

$$k_l = \max(k_{top}, \lceil \frac{L-l}{L} s \rceil)$$

This strategy increases the value of k in lower layers, widening the range of the selection at these points. At the top of the network, just before the dense layer, the result of this stack of layers is a feature graph that links sentence elements that were selected by being relevant to the classification task.

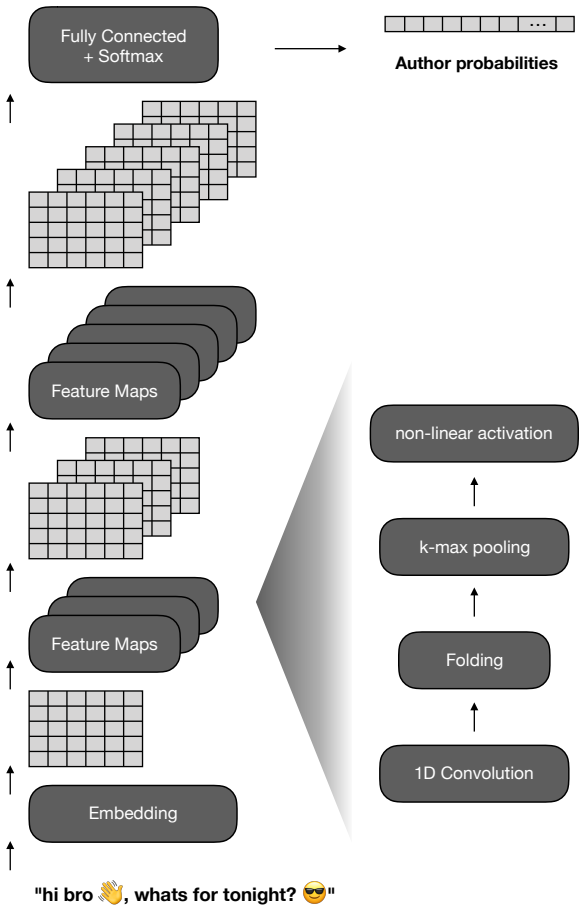

Fig. 1. Convolutional neural network model architecture applied for the problem of authorship attribution.

The Python code developed to implement this model is freely available for download and use [1].

4. EXPERIMENTAL RESULTS

4.1. Dataset

In order to train and evaluate our model, we downloaded 128 million messages (tweets) from 50,000 Twitter® users during April 2018. As a means to improve the quality of the dataset, we proceeded with some preprocessing and sanitization steps over it. Firstly, we restricted our data to English-speaker users through a parameter in Twitter® API and the application of Python langid library. Thereafter, we tried to identify bot profiles to be excluded from our dataset based on gestalt pattern matching, as bots could artificially make our task easier. We also filtered out tweets with less than three tokens since these samples do not provide meaningful information about the author and end up introducing noise into the classification task [18].

[1] https://github.com/theocjr/social-media-forensics/releases/tag/v1.0-icassp_2019

We have taken two additional steps in order to improve the quality of our data, trying to avoid user profiles that contain messages written by more than one person. The first step was to keep away from popular profiles, such as celebrities or politicians while the second was to exclude, from our dataset, messages classified as retweets. The latter was performed checking Twitter® metadata and also removing the ones that contain the pattern *RT @user*, indicating to be a retweet not identified by the social network API. Finally, as done by Rocha et al. [18], we replaced numbers, URLs, dates, times, hashtags and user references by specific tags in order to improve the classification process and also do not create artificial biases.

In the end, 70% of the data was employed for training the model with the remaining used for validation and testing purposes.

Due to restrictions of Twitter® API, we can not explicitly provide the dataset used in our experiments; however, the developed Python code used to download it directly from Twitter® is freely available [2].

4.2. Experimental Setup

Given the size of the hyper-parameter search space, we restricted our experiments to the scenario of classification of 50 authors with more than 2,000 messages each. With this strategy in hand, we were able to exploit more hyper-parameter sets, a crucial step for training this type of neural networks. We also employed dropout and L2 regularization methods to avoid overfitting.

We performed grid-search with the following set of hyper-parameters: learning rate, embedding dimension, size of convolutional filters, number of convolutional filters and k_{top} parameter (from the k-max pooling operation). Table 1 presents the values we used in this search, resulting in 2,187 different experiments.

Hyper-parameter	Tested values
learning rate	[0.01 , **0.1** , 0.5]
embedding dimension	[24 , **48** , 72]
convolutional layer 1 filter size	[**5** , 10 , 30]
convolutional layer 2 filter size	[**4** , 7 , 21]
convolutional layer 1 # filters	[**3** , 6 , 18]
convolutional layer 2 # filters	[6 , 12 , **36**]
k_{top}	[**2** , 4 , 12]

Table 1. Hyper-parameter values used in the grid-search and the best scenario found (in bold).

To conduct these experiments, we used a machine with 72 x Intel(R) Xeon CPU 2.30GHz cores, 512 GB of RAM and 6 x NVIDIA GEFORCE GTX 1080 Ti GPU cards. The training of the network does not demand such a powerful machine since it uses less than 500MB of RAM, however, a stronger infrastructure allows the deployment and assessment of different possible setups in parallel.

4.3. Results and Discussion

Each experiment was performed through 10 epochs, and the results for the best hyper-parameters configuration trained for 400 epochs can be found in Figure 2.

The baseline accuracy of 67.5% was generated through the application of Rocha et al. method [18] over our dataset since the accuracy presented in their paper (63.97%) was achieved over a different set of authors.

[2] https://github.com/theocjr/twitter-reader

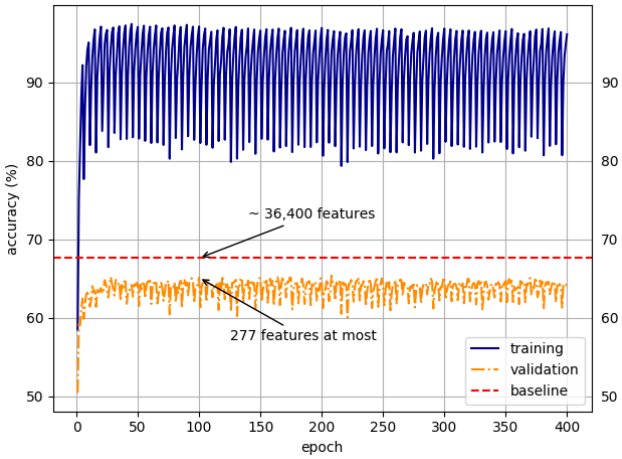

Fig. 2. Baseline, training and validation accuracies using the best hyper-parameters setting found.

It can be seen that the network achieved a training accuracy of 97.5% and a validation accuracy of 65%, right just below the baseline, almost reaching the 63% plateau already in the 5th epoch.

This validation result along with the high training accuracy reveal the power of this kind of model for the authorship attribution task, being able to identify author-specific stylometrics and important features like onomatopoeias, typically employed in social media parlance. We also applied baseline and our method over a test set, achieving respectively accuracies of 65% and 60%, confirming its potential.

Two additional advantages of our approach worth to mention are the absence of a hard human feature engineering step and the massive reduction in the input space dimensionality, decreasing from 36,400 dimensions [18] to less than 277 (at most when the tweet has 280 characters). This is a remarkable accomplishment, paving the way for different improvements in the future since it allows us to work with more training data, more suspect authors, using less memory and also to deal with the open-set scenario.

5. CONCLUSIONS AND FUTURE WORK

In this work, we showed a new approach to address the challenging problem of authorship attribution of micro-messages, by framing it as a signal processing problem. Our approach models small texts – messages from Twitter® social media network – as unidimensional signals and employs a deep convolutional neural network for feature representation and classification of the signal, that is, the attribution of a micro-message to an author. The proposed approach achieved promising results, closing to the current state of the art for the problem without heavy human feature engineering and offering a significant reduction in the input dimensionality.

Treating the text as a unidimensional signal to be fed into an appropriate convolutional neural network is indeed a promising path. Hence, we plan to perform many alternatives for improvement in future works, such as using deeper models allied with techniques for data augmentation.

Finally, we believe that our approach has the potential of coming close to the solution to this challenging and important problem of our modern society.

References

[1] Statista The portal for statistics. Number of monthly active twitter users worldwide. https://bit.ly/2dt7OI9, 2018. [Online; accessed on October 29th, 2018].

[2] Statista The portal for statistics. Number of monthly active facebook users worldwide. https://bit.ly/2daz7Yr, 2018. [Online; accessed on October 29th, 2018].

[3] Adrian Chen. The agency. https://nyti.ms/2rbKM0v, 2015. [Online; accessed on October 29th, 2018].

[4] Joseba Elola. Rebelião contra as redes sociais. https://bit.ly/2BD46sv, 2018. [Online; accessed on October 29th, 2018].

[5] Scott Shane. Mystery of russian fake on facebook solved, by a brazilian. https://nyti.ms/2x1LVbv, 2017. [Online; accessed on October 29th, 2018].

[6] Florence Davey-Attlee and Isa Soares. The fake news machine. inside a town gearing up for 2020. https://cnnmon.ie/2CTCgcM, 2017. [Online; accessed on October 29th, 2018].

[7] Nicholas Confessore, Gabriel J. X. Dance, Richard Harris, and Mark Hansen. The follower factory. https://nyti.ms/2rJ8YZM, 2018. [Online; accessed on October 29th, 2018].

[8] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*, 2017.

[9] Alex Heath. Facebook quietly updated two key numbers about its user base. https://read.bi/2pYvbPG, 2017. [Online; accessed on October 29th, 2018].

[10] Kasey Panetta. Gartner top strategic predictions for 2018 and beyond. https://gtnr.it/2ljsDMv, 2017. [Online; accessed on October 29th, 2018].

[11] Issie Lapowsky. Trust in social media withers in the industry's own backyard. https://bit.ly/2ErH0ri, 2018. [Online; accessed on October 29th, 2018].

[12] John F. Burrows. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and linguistic Computing*, 2(2):61–70, 1987.

[13] Andrew Queen Morton. *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. Simon & Schuster, 1978. ISBN 9780684155166.

[14] Patrick Juola. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334, 2008.

[15] Matthew L. Jockers and Daniela M. Witten. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, page fqq001, 2010.

[16] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60 (1):9–26, 2009.

[17] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[18] Anderson Rocha, Walter Scheirer, Christopher Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne Carvalho, and Efstathios Stamatatos. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33, 2017.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Intl. Conference on Machine Learning (ICML)*, pages 173–182, 2016.

[21] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research (JAIR)*, 57:345–420, 2016.

[22] Davide Castelvecchi. Deep learning boosts google translate tool. https://go.nature.com/2CO7alY, 2016. [Online; accessed on October 29th, 2018].

[23] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[24] Rie Johnson and Tong Zhang. Effective use of word order for text categorization with convolutional neural networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112, 2015.

[25] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 655–665, 2014.

[26] Yoon Kim. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[27] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[29] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 12(Aug):2493–2537, 2011.